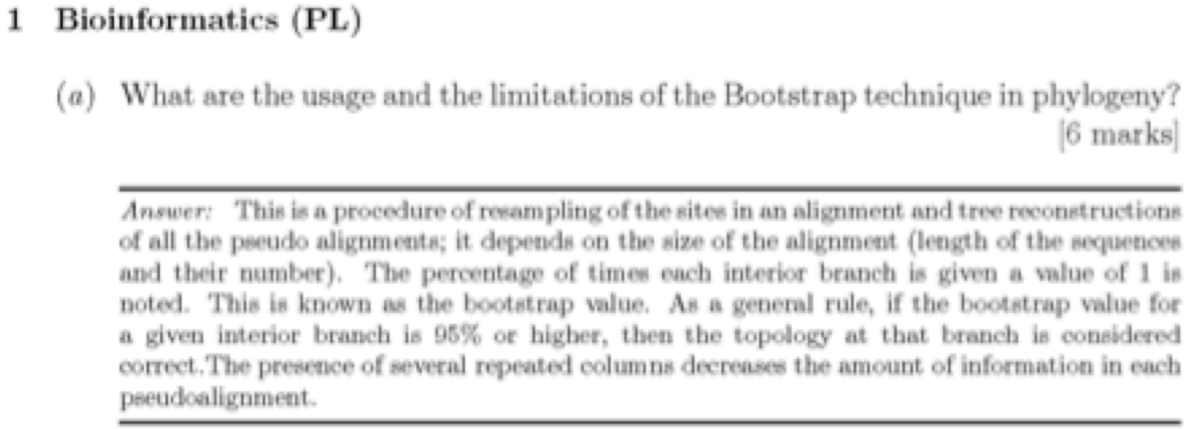
Some past exam questions

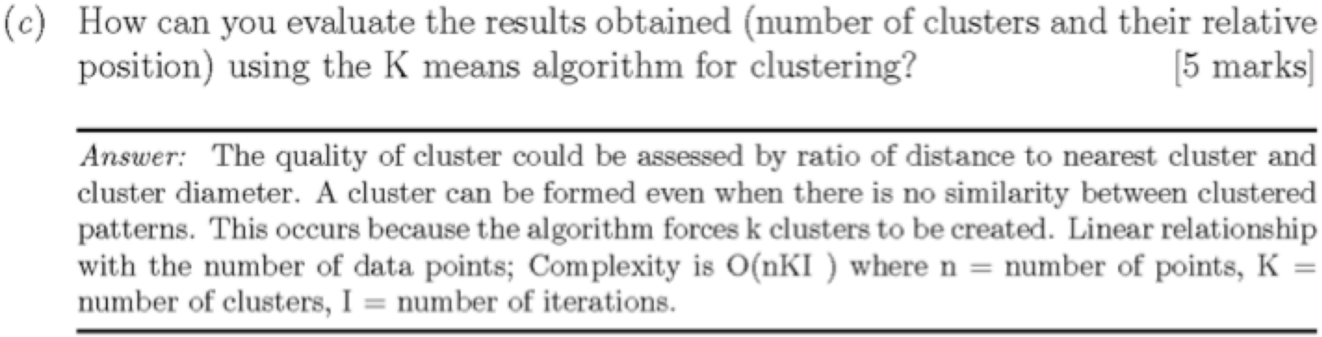
Exam questions



1 Bioinformatics (PL)

(*a*) What are the usage and the limitations of the Bootstrap technique in phylogeny? [6 marks]

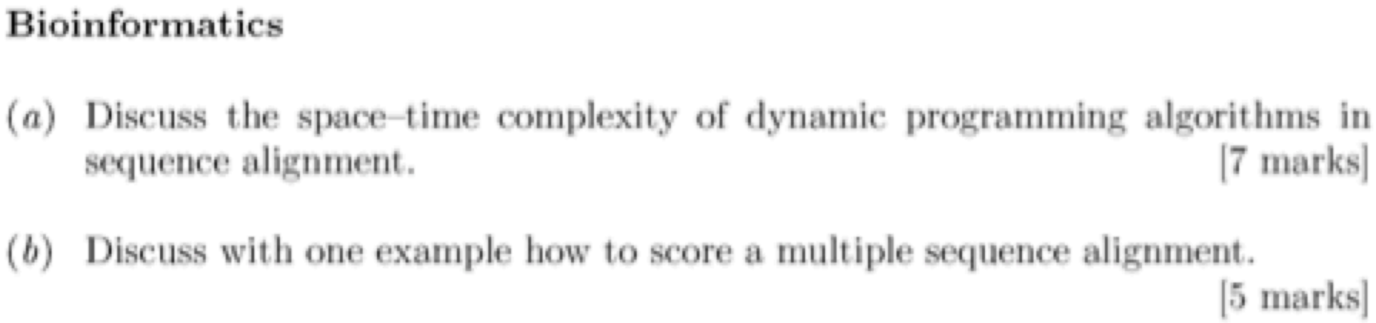
Answer: This is a procedure of resampling of the sites in an alignment and tree reconstructions of all the pseudo alignments; it depends on the size of the alignment (length of the sequences and their number). The percentage of times each interior branch is given a value of 1 is noted. This is known as the bootstrap value. As a general rule, if the bootstrap value for a given interior branch is 95% or higher, then the topology at that branch is considered correct.The presence of several repeated columns decreases the amount of information in each pseudoalignment.



(*c*) How can you evaluate the results obtained (number of clusters and their relative position) using the K means algorithm for clustering? [5 marks]

Answer: The quality of cluster could be assessed by ratio of distance to nearest cluster and cluster diameter. A cluster can be formed even when there is no similarity between clustered patterns. This occurs because the algorithm forces k clusters to be created. Linear relationship with the number of data points; Complexity is O(nKI) where n = number of points, K = number of clusters, I = number of iterations.

Exam questions



Bioinformatics

(*a*) Discuss the space–time complexity of dynamic programming algorithms in sequence alignment. [7 marks]

(*b*) Discuss with one example how to score a multiple sequence alignment. [5 marks]

Exam questions

1. Give the alignment matrix of the sequences `AATCGCGCGGT' and `ATGCGCCGT' assuming the following costs: Cost(a,a)=0; Cost(a,b)=3 when a ≠ b, Cost(a,-)=Cost(-,a)=2.
2. How would you set the function Cost in order to compute the longest subsequence common to x and y?
3. Describe the differences between the algorithms for global and local alignments
4. Which of the following reasons would lead you to use the Smith-Waterman local alignment algorithm instead of the Needleman-Wunsch global alignment algorithm?

Select all appropriate answers.

(a) Computer memory is too limited to compute the optimal global alignment.

(b) One wants to identify common protein domains in the two sequences.

(c) The sequences have very different lengths.

(d) Smith-Waterman is faster than Needleman-Wunsch on long sequences.

5. Describe the notion of a parsimonious phylogeny for a finite set of sequences and the hypothesis assumed on them

3 Bioinformatics (PL)

Given the two DNA sequences: GCACTT and CCCAAT

(*a*) Compute the alignment (using the edit graph) and the final score with the following rules: match score = +1, mismatch = −1, gap penalty = −1. [4 marks]

(*b*) Discuss how the alignment score and the quality of the result depend on the match score, mismatch, and gap penalty. [6 marks]

(*c*) Generate four, short DNA sequences (a,b,c,d) such that their relations as a tree are approximately the following: ((a,b),(c,d)). [5 marks]

(*d*) How is the score matrix used in phylogenetic tree building techniques? [5 marks]

1 Bioinformatics (PL)

(*a*) What are the usage and the limitations of the Bootstrap technique in phylogeny? [6 marks]

(*b*) We often use Hidden Markov Models (HMM) to predict a pattern (for instance the exons). How can you compute the number of True Positives, True Negatives, False Positives and False Negatives and use them to evaluate your HMM? [6 marks]

(*c*) How can you evaluate the results obtained (number of clusters and their relative position) using the K means algorithm for clustering? [5 marks]

HMM

(*b*) We often use Hidden Markov Models (HMM) to predict a pattern (for instance the exons). How can you compute the number of True Positives, True Negatives, False Positives and False Negatives and use them to evaluate your HMM?

[6 marks]

Answer:

(*i*) be predicted to occur: Predicted Positive (PP)

(*ii*) be predicted not to occur: Predicted Negative (PN)

(*iii*) actually occur: Actual Positive (AP)

(*iv*) actually not occur: Actual Negative (AN)

(*v*) True Positive $TP = PP \bigcap AP$

(*vi*) True Negative $TN = PN \bigcap AN$

(*vii*) False Negative $FN = PN \bigcap AP$

(*viii*) False Positive $FP = PP \bigcap AN$

(*ix*) Sensitivity: probability of correctly predicting a positive example Sn = TP/(TP + FN)

(*x*) Specificity: probability of correctly predicting a negative example Sp = TN/(TN + FP) or

(*xi*) probability that positive prediction is correct Sp = TP/(TP + FP)

3 Bioinformatics (PL)

Given the two DNA sequences: GCACTT and CCCAAT

(*a*) Compute the alignment (using the edit graph) and the final score with the following rules: match score = +1, mismatch = −1, gap penalty = −1. [4 marks]

(*b*) Discuss how the alignment score and the quality of the result depend on the match score, mismatch, and gap penalty. [6 marks]

(*c*) Generate four, short DNA sequences (a,b,c,d) such that their relations as a tree are approximately the following: ((a,b),(c,d)). [5 marks]

(*d*) How is the score matrix used in phylogenetic tree building techniques? [5 marks]

1 Bioinformatics (PL)

(*a*) What are the usage and the limitations of the Bootstrap technique in phylogeny? [6 marks]

(*b*) We often use Hidden Markov Models (HMM) to predict a pattern (for instance the exons). How can you compute the number of True Positives, True Negatives, False Positives and False Negatives and use them to evaluate your HMM? [6 marks]

(*c*) How can you evaluate the results obtained (number of clusters and their relative position) using the K means algorithm for clustering? [5 marks]

(*d*) What is the difference between the adjacency list and the accessibility list? [3 marks]

2009 Paper 9 Question 3

Bioinformatics

(*a*) Compute the global alignment between the two strings s1 = ACCGTT and s2 = AGTTCA, considering the following scoring parameters: +1 for match, −1 for mismatch, and −1 for a gap.

(*i*) What is the maximum similarity score between the two sequences s1 and s2? [2 marks]

(*ii*) Find an alignment with this similarity score. [2 marks]

(*iii*) Is the alignment you found unique, or are there multiple alignments achieving the maximum similarity score? [1 mark]

(*b*) Discuss the complexity of the Sankoff parsimony algorithm. [4 marks]

(*c*) Discuss the main differences between K-means, [redacted] and Markov clustering algorithms. [7 marks]

(*d*) Discuss the utility of the Gillespie algorithm in system biology. [4 marks]

2009 Paper 7 Question 5

Bioinformatics

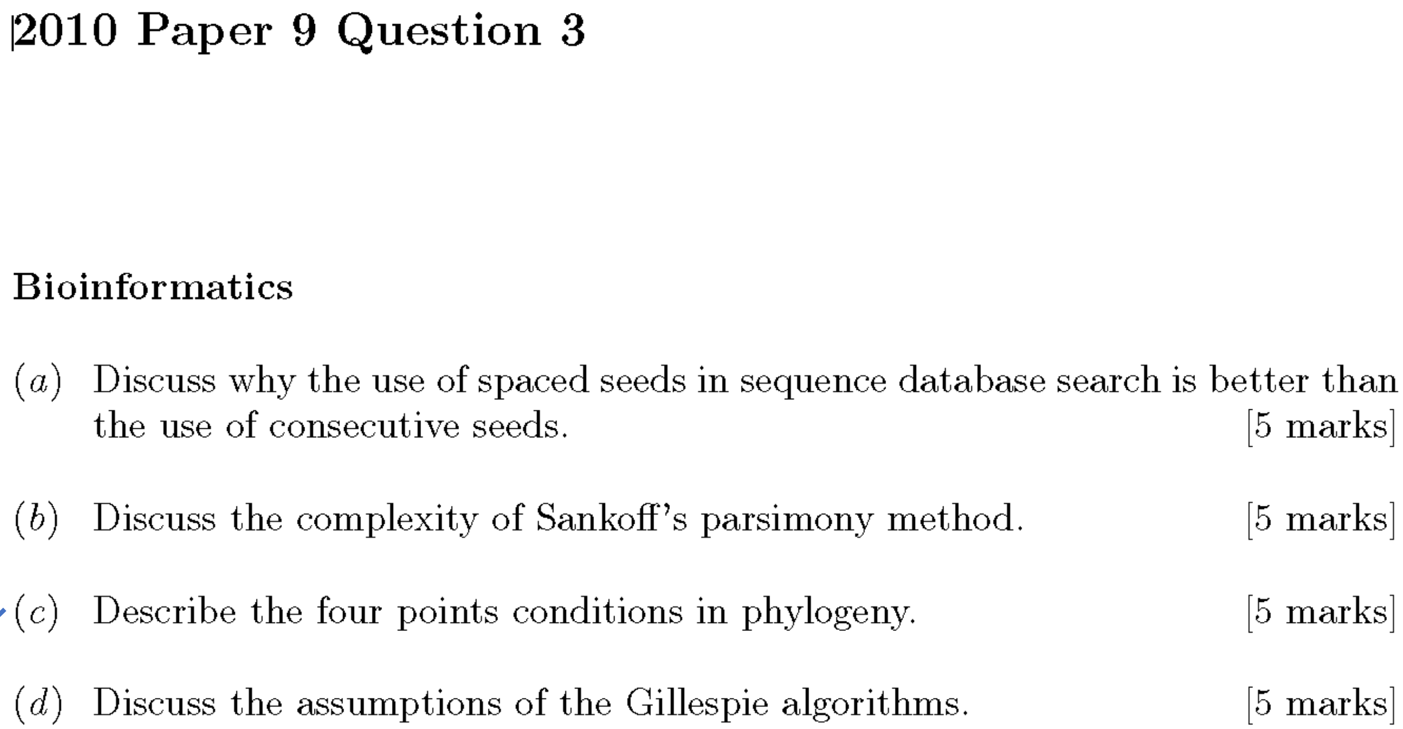
(*a*) Discuss, with one example, the complexity of the Nussinov algorithm for RNA folding. [5 marks]

(*b*) In the context of algorithms on strings, what is the advantage of using spaced seeds in database search? [3 marks]

(*c*) Hidden Markov models (HMM) are used to identify genes in genome sequencing projects.

(*i*) Describe how you would build a hidden Markov model to identify genes in a genome sequence. [7 marks]

(*ii*) How would you assess the sensitivity and specificity performance of the HMM? [5 marks]



2010 Paper 9 Question 3

Bioinformatics

(*a*) Discuss why the use of spaced seeds in sequence database search is better than the use of consecutive seeds. [5 marks]

(*b*) Discuss the complexity of Sankoff's parsimony method. [5 marks]

(*c*) Describe the four points conditions in phylogeny. [5 marks]

(*d*) Discuss the assumptions of the Gillespie algorithms. [5 marks]

c) Describe the additivity condition in phylogeny (the course content has changed)

2010 Paper 7 Question 5

Bioinformatics

(*a*) Discuss the space–time complexity of dynamic programming algorithms in sequence alignment. [7 marks]

(*b*) Discuss with one example how to score a multiple sequence alignment. [5 marks]

1 Bioinformatics (PL)

(*a*) Considerable recent research has focused on *alignment of sequences*.

(*i*) Why do we use dynamic programming algorithms for sequence alignment problems? [3 marks]

(*ii*) Describe what needs to be taken into account for gaps in DNA sequence alignment. [3 marks]

(*b*) Considerable recent research has focused on sequence database search methods.

(*i*) What are the most important differences between PatternHunter, BLAST, Smith-Waterman and Needleman-Wunsch algorithms? [9 marks]

(*ii*) Compare the heuristic used by Clustal with a dynamic programming algorithm for multiple alignment. [5 marks]

3 Bioinformatics (PL)

(*a*) Considerable recent Bioinformatics research has focused on *phylogenetics*.

(*i*) What is the motivation for this work? [1 mark]

(*ii*) Describe with the aid of examples *two* different techniques for phylogeny. In each case discuss the issues of complexity and performance. [4 marks each]

(*b*) Considerable recent Bioinformatics research has focused on *structure prediction from sequence data*.

(*i*) Describe how you would build a hidden Markov model (HMM) to identify membrane segments in aminoacid sequences. [6 marks]

(*ii*) How you would assess the sensitivity and specificity performance of your HMM? [5 marks]

2009 Paper 9 Question 3

Bioinformatics

(*a*) Compute the global alignment between the two strings s1 = ACCGTT and s2 = AGTTCA, considering the following scoring parameters: +1 for match, −1 for mismatch, and −1 for a gap.

(*i*) What is the maximum similarity score between the two sequences s1 and s2? [2 marks]

(*ii*) Find an alignment with this similarity score. [2 marks]

(*iii*) Is the alignment you found unique, or are there multiple alignments achieving the maximum similarity score? [1 mark]

(*b*) Discuss the complexity of the Sankoff parsimony algorithm. [4 marks]

(*c*) Discuss the main differences between K-means, [redacted] and Markov clustering algorithms. [7 marks]

(*d*) Discuss the utility of the Gillespie algorithm in system biology. [4 marks]

2010 Paper 7 Question 5

Bioinformatics

(*a*) Discuss the space–time complexity of dynamic programming algorithms in sequence alignment. [7 marks]

(*b*) Discuss with one example how to score a multiple sequence alignment. [5 marks]

1 Bioinformatics (PL)

(*a*) What are the usage and the limitations of the Bootstrap technique in phylogeny? [6 marks]

(*b*) We often use Hidden Markov Models (HMM) to predict a pattern (for instance the exons). How can you compute the number of True Positives, True Negatives, False Positives and False Negatives and use them to evaluate your HMM? [6 marks]

(*c*) How can you evaluate the results obtained (number of clusters and their relative position) using the K means algorithm for clustering? [5 marks]

(*d*) What is the difference between the adjacency list and the accessibility list? [3 marks]

COMPUTER SCIENCE TRIPOS Part II – 2013 – Paper 9

1 Bioinformatics (PL)

phylogeny

(*a*) What are the usage and the limitations of the Bootstrap technique in phylogeny? [6 marks]

Answer: This is a procedure of resampling of the sites in an alignment and tree reconstructions of all the pseudo alignments; it depends on the size of the alignment (length of the sequences and their number). The percentage of times each interior branch is given a value of 1 is noted. This is known as the bootstrap value. As a general rule, if the bootstrap value for a given interior branch is 95% or higher, then the topology at that branch is considered correct.The presence of several repeated columns decreases the amount of information in each pseudoalignment.

HMM

(*b*) We often use Hidden Markov Models (HMM) to predict a pattern (for instance the exons). How can you compute the number of True Positives, True Negatives, False Positives and False Negatives and use them to evaluate your HMM?

[6 marks]

Answer:

(*i*) be predicted to occur: Predicted Positive (PP)

(*ii*) be predicted not to occur: Predicted Negative (PN)

(*iii*) actually occur: Actual Positive (AP)

(*iv*) actually not occur: Actual Negative (AN)

(*v*) True Positive $TP = PP \bigcap AP$

(*vi*) True Negative $TN = PN \bigcap AN$

(*vii*) False Negative $FN = PN \bigcap AP$

(*viii*) False Positive $FP = PP \bigcap AN$

(*ix*) Sensitivity: probability of correctly predicting a positive example Sn = TP/(TP + FN)

(*x*) Specificity: probability of correctly predicting a negative example Sp = TN/(TN + FP)
or

(*xi*) probability that positive prediction is correct Sp = TP/(TP + FP)

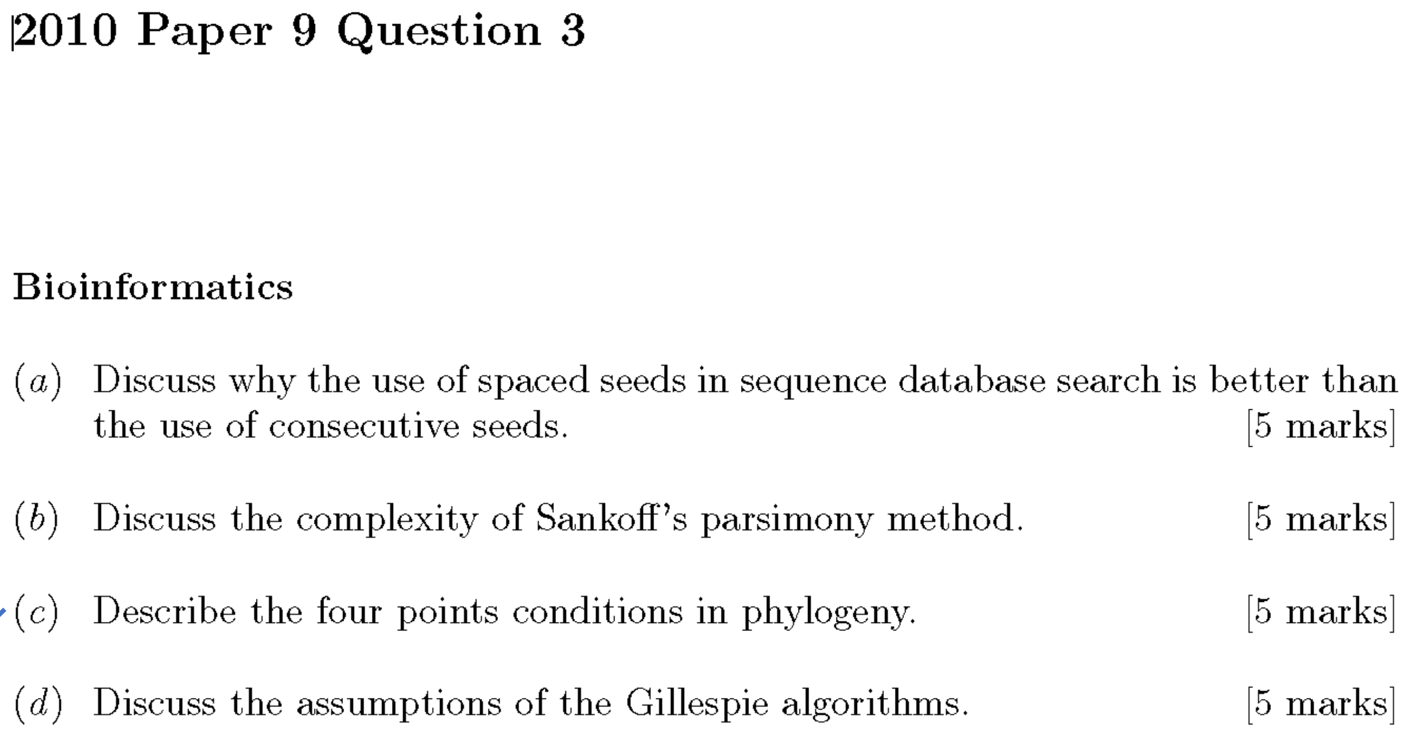
---

clustering

(*c*) How can you evaluate the results obtained (number of clusters and their relative position) using the K means algorithm for clustering? [5 marks]

Answer: The quality of cluster could be assessed by ratio of distance to nearest cluster and cluster diameter. A cluster can be formed even when there is no similarity between clustered patterns. This occurs because the algorithm forces k clusters to be created. Linear relationship with the number of data points; Complexity is O(nKI) where n = number of points, K = number of clusters, I = number of iterations.

Answer: The adjacency list is the set of nodes (genes) adjacent to (directly influenced by) node under consideration. One might also call it the list of nearest neighbors in the gene network, or the list of direct regulatory interactions. The accessibility list is the list of perturbation effects or the list of regulatory effects. Acc(i) is the set of nodes that can be reached from node i by following all paths of directed edges leaving i.



2010 Paper 9 Question 3

Bioinformatics

(*a*) Discuss why the use of spaced seeds in sequence database search is better than the use of consecutive seeds. [5 marks]

(*b*) Discuss the complexity of Sankoff's parsimony method. [5 marks]

(*c*) Describe the four points conditions in phylogeny. [5 marks]

(*d*) Discuss the assumptions of the Gillespie algorithms. [5 marks]

c) Describe the additivity condition in phylogeny

COMPUTER SCIENCE TRIPOS Part II – 2012 – Paper 9

1 Bioinformatics (PL)

(*a*) Considerable recent research has focused on *alignment of sequences.*

(*i*) Why do we use dynamic programming algorithms for sequence alignment problems? [3 marks]

(*ii*) Describe what needs to be taken into account for gaps in DNA sequence alignment. [3 marks]

(*b*) Considerable recent research has focused on sequence database search methods.

(*i*) What are the most important differences between PatternHunter, BLAST, Smith-Waterman and Needleman-Wunsch algorithms? [9 marks]

(*ii*) Compare the heuristic used by Clustal with a dynamic programming algorithm for multiple alignment. [5 marks]

COMPUTER SCIENCE TRIPOS Part II – 2012 – Paper 7

3 Bioinformatics (PL)

(*a*) Considerable recent Bioinformatics research has focused on *phylogenetics*.

(*i*) What is the motivation for this work? [1 mark]

(*ii*) Describe with the aid of examples *two* different techniques for phylogeny. In each case discuss the issues of complexity and performance. [4 marks each]

(*b*) Considerable recent Bioinformatics research has focused on *structure prediction from sequence data*.

(*i*) Describe how you would build a hidden Markov model (HMM) to identify membrane segments in aminoacid sequences. [6 marks]

(*ii*) How you would assess the sensitivity and specificity performance of your HMM? [5 marks]